UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4900/9900 — Statistical methods and applications
Day of examination:	June 13th, 2022
Examination hours:	15.00 – 19:00

This problem set consists of 8 pages.

Appendices:	Tables for normal, t-, χ^2- and F-distributions
Permitted aids:	All printed, hand-written resources. Approved calculator.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

From a study on forest fires in the Montesinho Natural Park by Cortez & Morais (2007), we select the subset of the observations for which the fire caused a burned area (area > 0). For these 270 observations, we have information about:

- area: total area in ha burned by the fire;
- season: season of the year (4 categories);
- wind: wind speed in km/h;
- FFMC: Fine Fuel Moisture Code from FWI;
- DMC: Duff Moisture Code from FWI;
- DC: Drought Code from FWI;
- ISI: Initial Spread Index from FWI;
- temp: temperature in Celsius degrees;
- RH: relative humidity in %;
- wind: wind speed in km/h;
- rain: outside rain in mm/m2.

All 4 codes from the FWI (Fire Weather Index) are continuous variables. The response variable is area.

(Continued on page 2.)

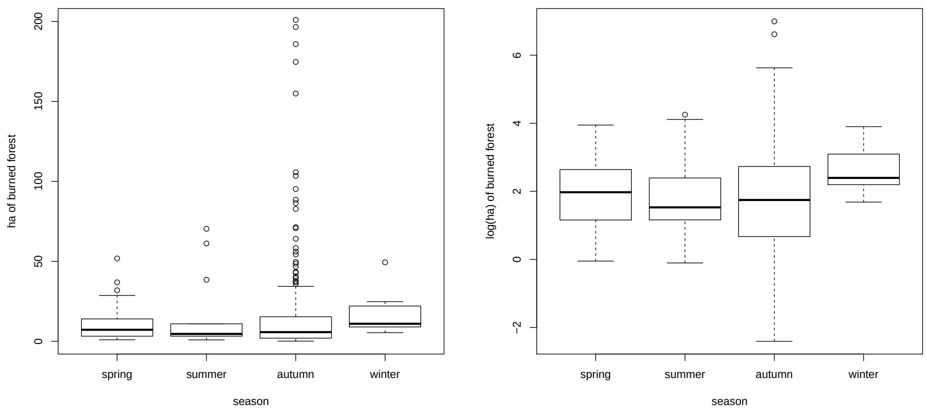

Figure 1: Area burned for each season of the year. Left: original scale; right: logarithm scale.

a

The first goal is to evaluate if there is any statistically significant difference in the size of the fires (understood as the forest area burned) among the four seasons. The goal is to perform an ANOVA to test the hypothesis of equal means. Before performing the analysis, we need to decide if we want to logarithmically transform the response variable. Looking at Figure 1 (here above), briefly describe what a box-plot is and explain why in this case it seems a good idea to take the logarithmic transformation of the response.

b

It is now time to perform the test. Write down formally the null and the alternative hypothesis for the test mentioned in point (a) and use the following information,

- model sum of square: 8.7;
- residual sum of square: 618.3;

to perform the test. Report the degrees of freedom, the value of the test-statistic, and the p-value. Comment on the result.

NB: as you need to read the values of the distribution from a table (the tables are in the Appendix), here, and in the rest of the exam, it is sufficient to provide an “approximation” (as much as you can get from the tables) for the p-value, it is not required the precise number.

c

Let us now fit a linear regression model. Starting from the model only containing the information about the season, we aim at performing one step of “forward selection”. Consider the following R output (see next page):

(Continued on page 3.)

```
Single term additions

Model:
log(area) ~ season
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>               618.32 231.72
FFMC    1    0.0027 618.32 233.72  0.0012 0.97290
DMC     1    9.0806 609.24 229.72  3.9498 0.04791
DC      1    1.6120 616.71 233.01  0.6927 0.40600
ISI     1    3.6012 614.72 232.14  1.5524 0.21388
temp    1    0.9213 617.40 233.32  0.3954 0.52999
RH      1    1.2385 617.08 233.18  0.5319 0.46647
wind    1    0.2270 618.10 233.62  0.0973 0.75529
rain    1    0.1444 618.18 233.66  0.0619 0.80372
```

Briefly explain the “forward selection” procedure for model selection.
Then identify the variable that should be added to the model in the first iteration and explain the reason behind your choice.

d

When the variable is added to the model, we obtain the following (modified) R output, where X should be replaced with the name of the variable selected at point (c).

```
Call:
lm(formula = log(area) ~ season + X, data = data)

Coefficients:
                Estimate Std. Error
(Intercept)     1.885746   0.286991
season2_summer -0.215159   0.509637
season3_autumn -0.646478   0.368515
season4_winter  0.536254   0.493514
X               0.003958   0.001992

Residual standard error: 1.516 on 265 degrees of freedom
Multiple R-squared:  0.02832,   Adjusted R-squared:  0.01365
F-statistic: 1.931 on 4 and 265 DF,  p-value: 0.1056
```

Consider now the meaning of the regression coefficients. When the value of the variable X increases by 10, how does the expected size of burned area change? (Hint: remember that we are modelling its logarithm)
Moreover, provide the expected size of burned area for a fire that happens in summer and for a value of the variable X equal to 50.

e

As the model fitted at point (d) is not really satisfactory, we try to add all available variables to the model. Here the result:

(Continued on page 4.)

```
Call:
lm(formula = I(log(area)) ~ season + FFMC + DMC + DC + ISI +
    temp + RH + wind + rain, data = data)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.9257     3.6820   0.251   0.8017
season2_summer    -0.1693     0.5480  -0.309   0.7577
season3_autumn    -0.5006     0.6537  -0.766   0.4445
season4_winter     0.3992     0.6476   0.616   0.5381
FFMC               0.0153     0.0413   0.371   0.7108
DMC                0.0042     0.0023   1.795   0.0739
DC                -0.0001     0.0001  -0.094   0.9256
ISI               -0.0555     0.0347  -1.600   0.1109
temp               0.0049     0.0279   0.174   0.8618
RH                -0.0066     0.0086  -0.768   0.4434
wind               0.0447     0.0572   0.782   0.4348
rain               0.0555     0.2400   0.231   0.8173
---

Residual standard error: 1.524 on 258 degrees of freedom
Multiple R-squared:  0.0441,   Adjusted R-squared:  0.003347
F-statistic: 1.082 on 11 and 258 DF,  p-value: 0.3759
```

Explain why the model fitted at point (d) is not really satisfactory and, by looking at the outcome above, guess whether the new model is an improvement with respect to the one obtained at point (d). Justify your opinion.
Perform a test to check your guess, i.e., whether this last model or that fitted at point (d) is preferable. You may want to use the results of the following R output (remember that X is actually one of the variables of the last model),

```
Analysis of Variance Table

Model 1: log(area) ~ season + X
Model 2: log(area) ~ season + FFMC + DMC + DC + ISI + temp + RH +
 wind + rain
  Res.Df    RSS
1    265 609.24
2    258 599.35
```

Problem 2

The UCI Machine Learning Repository contains a dataset by Guvenir et al. (1998) about arrhythmia. An arrhythmia is defined as an abnormality of the heart's rhythm. In this exercise, we focus on the clinical information included in the data, ignoring the ECG measurements. As a consequence, we have 5 variables,

- `Age`: Age in years;

(Continued on page 5.)

- `Sex`: Sex (0 = male; 1 = female);
- `Height`: Height in centimetres;
- `Weight`: Weight in kilograms;
- `y`: presence of arrhythmia (0: no, 1: yes).

a

Imagining that the sample collected in the study is representative of the population, perform a test to check if the proportion of people with arrhythmia is the same in the male and female population. In the sample,

```
        normal arrythmia Sum
male        85       117 202
female     160        88 248
Sum        245       205 450
```

After having briefly described the concepts of “excess risk”, “relative risk” and “odds ratio”, compute them for the male versus female population with these data.

b

Consider now the people’s age. By fitting a logistic regression model, we obtain the following R output (edited):

```
Call:
glm(formula = y ~ Age, family = binomial, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.335629   0.289605  -1.159    0.246
Age          0.003369   0.005853     ...      ...

    Null deviance: 620.27  on 449  degrees of freedom
Residual deviance: 619.94  on 448  degrees of freedom
```

Provide the interpretation for the regression coefficient related to the variable Age. Compute the z-value and the p-value related to it and provide the 95% confidence interval.

c

Also in this case, we also fit a model with all the explanatory variables we have available. The resulting model is (in a modified R output)

```
Call:
glm(formula = y ~ Age + Sex + Height + Weight, family = binomial,
    data = data)
```

(Continued on page 6.)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.847803   2.228116   3.073  0.00212 **
Age          0.006817   0.006559   1.039  0.29865
Sexfemale   -1.335153   0.243948  -5.473 4.42e-08 ***
Height      -0.039511   0.014137  -2.795  0.00519 **
Weight      -0.001999   0.007700  -0.260  0.79520

    Null deviance: 620.27  on 449  degrees of freedom
Residual deviance: 585.36  on 445  degrees of freedom
```

Comment the fitted model, in particular discussing:

- what does the regression coefficient of Sex mean here?
- is its significance level in line with the results of point (a)?
- why did the regression coefficient of the variable Age change (although minimally) with respect to that obtained at the point (c)?
- does this model give a better fit than the one of point (c)? Perform a test to support your answer.

d

Finally, consider the following R output for a model that also includes a second order effect for Age,

```
Call:
glm(formula = y ~ Age + I(Age^2) + Sex + Height + Weight,
    family = binomial, data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.9060592  2.3973921   2.881  0.00397 **
Age         -0.0784407  0.0348305  -2.252  0.02432 *
I(Age^2)     0.0009048  0.0003630   2.493  0.01268 *
Sexfemale   -1.2326965  0.2539864  -4.853 1.21e-06 ***
Height      -0.0320293  0.0151363  -2.116  0.03434 *
Weight       0.0043771  0.0081598   0.536  0.59167

    Null deviance: 620.27  on 449  degrees of freedom
Residual deviance: 579.00  on 444  degrees of freedom
AIC: 591
```

Why are the regression coefficients related to Age significant now, in contrast to the result of point (c)? If one would have fitted a generalised additive model allowing a smooth function to capture the effect of Age, which form would have it taken? Draw it in a plot.

(Continued on page 7.)

Problem 3

Krall et al. (1975) discussed a small study about multiple myeloma, in which some potential prognostic factors (i.e., characteristics of a patient that are useful to estimate the chance of death) have been considered. From the original data, they selected the 65 patients with complete data, 48 of which died at the time of the study (events).

a

Based on the gender and the presence of a specific protein in the urine, we have created four groups of patients. To evaluate if the survival function was the same in the four group, a logrank test has been performed and the following (edited) R output obtained:

```
Call:
survdiff(formula = y ~ groups)

            N Observed Expected (O-E)^2/E (O-E)^2/V
Group 1    15       12    13.37     0.141     0.208
Group 2    23       18    14.57     0.809     1.240
Group 3     8        8    11.13     0.881     1.361
Group 4    19       10     8.93     0.128     0.180

Chisq = 2.4 on ? degrees of freedom, p = ?
```

State the null and the alternative hypothesis.
With the help of the tables provided in the Appendix, provide the p-value (and how you ended up with that number) and argue in favour or against the idea that there is support in the data for rejecting the null hypothesis.

b

Consider now the first group. The survival times in months have been reported in the following table:

Time	Dead/Alive	Time	Dead/Alive	Time	Dead/Alive
1	D	11	A	41	D
2	D	14	D	51	D
2	D	35	D	54	D
4	A	37	D	67	D
11	D	41	A	89	D

Use the Kaplan-Meier estimator to estimate the survival function and plot in a graph. Identify the median survival time and show in the plot how to find it graphically.

(Continued on page 8.)

c

Consider the following (edited) R output, obtained by fitting a Cox model to the data. Here `logBUN` measures the logarithm of the amount of urea nitrogen in the blood (in mmol per litre), `sex` codifies the gender (0 - male, 1 - female), `BJprotein` indicates whether the Bence Jone protein is present in the urine at diagnosis (1 - present, 2 - none), and `age` the patient's age at the beginning of the study (in years):

```
Call:
coxph(formula = y ~ logBUN + age + sex + BJprotein, data = X)

  n= 65, number of events= 48

                coef exp(coef) se(coef)      z Pr(>|z|)
logBUN       2.15209   8.60285  0.63593  3.384 0.000714
age         -0.02130   0.97892  0.01679 -1.269 0.204464
sex1        -0.07191   0.93062  0.31217 -0.230 0.817822
BJprotein2   0.66166   1.93801  0.33331  1.985 0.047127

           exp(coef) exp(-coef) lower .95 upper .95
logBUN        8.6028     0.1162    2.4736    29.919
age           0.9789     1.0215    0.9472     1.012
sex1          0.9306     1.0746    0.5047     1.716
BJprotein2    1.9380     0.5160    1.0084     3.724

Likelihood ratio test= 13.23  on ... df,   p=...
```

Interpret the model, providing the values and the meaning of all hazard ratios.
Identify the significant (at level $\alpha = 0.05$) prognostic factors and, only for those, report the 95% confidence interval. What can we say about the significance by only looking at these intervals?
Finally, compute the p-value for the likelihood ratio test and comment on the result.

References

CORTEZ, P. & MORAIS, A. D. J. R. (2007). A data mining approach to predict forest fires using meteorological data. In *Proceedings of the 13th Portuguese Conference on Artificial Intelligence.*

GUVENIR, H., ACAR, B. & MUDERRISOGLU, H. (1998). Arrhythmia. UCI Machine Learning Repository.

KRALL, J. M., UTHOFF, V. A. & HARLEY, J. B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics* **31**, 49–57.

THE END